Criticism

Volume 61 | Issue 3 Article 6

2019

Deforming Shakespeare's Sonnets: Topic Models as Poems

Amanda Henrichs
Amherst College, ahenrichs@amherst.edu

Deforming Shakespeare's Sonnets: Topic Models as Poems

Cover Page Footnote

Thanks to Peter Berek for his generous comments on a very early version of this essay; thanks also to Laura Estill and Eric Johnson for convening the workshop "Shakespeare by the Numbers" at the 2017 meeting of the Shakespeare Association of America, which sparked my interest in topic models as generative products. My sincere gratitude also goes to the anonymous reviewers of the essay for their constructive feedback.

Deforming Shakespeare's *Sonnets*: Topic Models as Poems

Deformance and Generative Criticism of Canonical Texts

In their essay "Deformance and Interpretation," Lisa Samuels and Jerome McGann follow Emily Dickinson's suggestion that "Reading Backward" could "overtake the mind": "Did you ever read one of her Poems backward, because the plunge from the front overturned you? I sometimes (often have, many times) have—a Something overtakes the Mind."[1] Dickinson was "overturned" by the usual approach to a poem; and so she read backward, with the result that, "a Something overtakes the Mind." There are mental and embodied consequences to reversing the reading procedure, and Samuels and McGann pursue reading backward as a path toward what they call "deformative criticism."[2] Ultimately, in being overtaken, the critic can move beyond interpretation, or the search for a singular meaning in the creative work. Instead, deformative criticism takes seriously the performative aspects of the poetic object; actually reading a poem backward illuminates the fact that interpretation is a secondary activity, based on the pre-existing "systemic intelligibility" of the poem.[3] Importantly, the poem's "systemic intelligibility" is not determined by the author of the work; because poems are not primarily communicative texts, there is no real need to maintain authorial intention. Samuels and McGann highlight, however, that most traditional literary criticism effectively reifies the poetic text:

> Critical and interpretive limits are thus regularly established … [and] taken as inviolable. From an interpretive point of view, this assumption brackets off from attention crucial features of imaginative works, features wherein the elemental forms of meaning are built and elaborated. These forms are so basic and conventionally governed—they are alphabetical and diacritical; they are the rules for character formation, character arrangement, and textual space, as well as for the

> structural forms of words, phrases, and higher morphemic and phonemic units—that readers tend to treat them as preinterpretive and precritical. In truth, however, they comprise the operating system of language, the basis that drives and supports the front-end software. That computing metaphor explains why most readers do not fool around with these levels of language. [4]

This passage asserts that critics take a poem's form as absolutely "inviolable," as a stable and permanent fact of the work. Samuels and McGann argue that as a result of this reification, critics thus "bracket off" the fact that language is made of letters, which themselves have shapes and normative appearances. These letters in turn form words, whose spelling is (for the most part) fixed; and set phrases and idiomatic expressions give structure to linguistic communication. The authors salvage words, phrases (etc.) from the obscurity of *technique*; they explain these supposedly "preinterpretive" aspects of linguistic creativity with a computational metaphor: letter-shapes are like a computer program; they are the "operating system of language" that "drives and supports the front-end software." What deformative criticism offers, then, is potential attention to the material medium of language, deploying a "computing metaphor" to explain why "most readers" do not pay attention to this medium. In this context, *most readers* are humanist literary scholars; and they do not pay attention to this "level of language" because it is, presumably, *just too hard* to do so. In fact, the recent rise of digital, experimental, and computational approaches to the humanities increase the opportunities available for critics to pay attention to language in this way, and to do so on an unprecendented scale.

This essay follows the computational thread Samuels and McGann begin to unspool and runs another highly-canonical text, Shakespeare's *Sonnets*, through a topic modeling

package in the statistical software R. While topic models are a common tool in digital humanities, the *Sonnets* have not been topic modeled. There are many good reasons for this, a few of which I will explain shortly; and yet there are good reasons to bring together a highly useful tool and this very canonical text, both to learn what the tool can offer and because close reading the products of distant reading demonstrates the socially-embedded nature of critical process.

As an author steeped in the humanist educational system of late sixteenth-century England, Shakespeare relies on the forms of his poetry to perform communicative functions. What unites humanist poems and digital humanities word clouds is an abiding concern with form, and particularly form as endowed with social meaning. Taken together, theories of early modern poetic form and modern digital humanities topic modeling practices emphasize that digital humanities products are not transparent keys to the text: they are generative, and are best when read like poems, a shaped remediation of language. And as generative products, they further illuminate the constructed nature of the processes behind both computational and traditional literary criticism.

Goals and Limitations of Computing Shakespeare

While traditional literary criticism is focused on the written product (the article, the chapter, the conference paper), both deformance and what Stephen Ramsay calls "algorithmic criticism" are focused on the critical process that precedes those products. Ramsay argues, like Samuels and McGann, that "Literary-critical interpretation is not just a qualitative matter; it is also an insistently subjective manner of engagement."[5] If literary criticism is deeply subjective in its choice of texts and its preference for certain approaches to those texts (i.e., psychoanalytic, New Formalist, New Historicist, queer theory, critical

race theory, and so on), it is also most often evaluated not according to some standard of truth, but rather "in terms of the nature and depth of the discussions that result."[6] Scholars evaluate ideas based on how much they make us think: the goal is never to end the discussion, but rather to deepen it, broaden it, to bring in multiple perspectives that enrich our understanding of the text in question. Ramsay extends this to computational approaches to literary study, claiming that "if text analysis is to participate in literary critical endeavor in some manner beyond fact-checking, it must endeavor to assist the critic in the unfolding of interpretative possibilities."[7] This is precisely what I propose to do: to transform the *Sonnets,* to deform them, in order to deepen the discussion and unfold further "interpretative possibilities."

While it is not explicitly an act of either deformative or algorithmic criticism, Stephen Booth's 1979 *Essay on Shakespeare's Sonnets* plumbs the depths of what Samuels and McGann would identify as part of the "operating system of language": not only the meaningful content words, but also the function words of the poems, are examined in a laborious human-powered process. These words, like *and, but,* or *the,* serve as the linguistic glue of Shakespeare's poetic project. Booth focuses on the function words in order to argue that Shakespeare's *Sonnets* had no single governing pattern; rather, the very multiplicity of patterns both within single sonnets and across the collection as a whole meant that no single pattern could take priority.[8] Indeed, the patternings of function words such as these have been taken up by computational humanist scholars in authorship arguments including but not limited to Shakespeare. Hugh Craig, Jonathan Hope, Michael Witmore, and Jonathan Lamb have all used computational tools such as word searches, stylometry, and sentiment

analysis on Shakespeare's plays, and Booth's argument thus set a precedent for similar computational approaches.

Given Shakespeare's uber-canonical status and the pre-existing interest in the nuts and bolts of the poet's language, there is surprisingly little work done on the plays via the popular topic modeling method. The only scholars to topic model the semantic content of the plays (as opposed to non-semantic features like character names) are Laura Estill and Luis Meneses, in a recent article on the Henriad.[9] They attribute the "relative lack of topic modeling" of the plays "to the fact that [Shakespeare's] most popular works today are plays; [and] rather than having an overarching narrative or authorial voice, Shakespeare dramatizes many characters with distinct voices."[10] In other words, because the plays shift between characters with varying and distinct voices, topic modeling algorithms would have a difficult time detecting latent patterns across the text; one of the things topic models assume is a unified authorial voice.

The *Sonnets* pose a different challenge. The diffuse voicing of the plays is less of a problem, but the available corpus dramatically shrinks, and topic modeling—or any computational approach—requires a large corpus, preferably on the order of hundreds of thousands (if not millions) of words. Craig points out that Shakespeare's dramatic corpus is proportionally larger than that of his contemporaries, which means that applying computational tools to the plays will produce more reliable results, mathematically speaking.[11] Topic modeling could thus be applied to plays, and the results could be regarded as significant. In contrast, because the non-dramatic corpus is so small, the results will not be statistically significant. Thus, though Shakespeare's non-dramatic poems do

present a more unified authorial voice, as a corpus they are arguably too small for computational techniques.[12]

This leads to the most important mathematical objection to topic modeling: it is explicitly developed as a tool for discovery in a corpus too large for human readers to successfully extract the subtle linguistic features that unify the text. David Blei, one of the original authors of LDA (or latent Dirichlet association, the most straightforward method for topic modeling) writes: "Imagine *searching and exploring* documents based on the themes that run through them"; and "Topic models are algorithms for *discovering* the main themes that pervade a large and otherwise unstructured collection of documents"; and that topic models "have been used to *find* patterns in genetic data, images, and social networks" (all emphases mine).[13] The importance of topic modeling as a digital humanist tool lies in its capacity for exploration and discovery. Topic models are meant to explore very large corpora, and since the *Sonnets* corpus is so small, we simply haven't needed to. The *Sonnets* are thus the delicate fossil that the topic modeling shovel is too big to excavate appropriately.

Taking the above objections into account, there are still two persuasive reasons to topic model the *Sonnets,* and particularly to output those topic models as word clouds. First, following Ramsay, doing so produces linguistic objects that themselves invite further interpretation and discussion. Second, and relatedly, word clouds seem more easily interpretable by literary scholars, which helps illuminate the constructed nature of even traditional literary criticism. Here, I want to argue for the value of applying our carefully-honed skills of close reading, not just to great texts, but also to the products of distant reading, of taking word clouds seriously as subjects of literary interpretation. Thereby a

computational deformance emerges, illuminated by how Shakespeare and his contemporaries viewed form as socially communicative.

Communicative Functions In and Through Form in Renaissance Poetry

The practice of shaping language is not unique to topic models nor to early modern poetry; indeed, poetry is at its heart a process of formalizing language in order to communicate in ways prohibited by the structures of prose. Mutlu Konuk Blasing argues for the defamiliarizing power of poetic form, suggesting that it is poetic form which not only shapes language into certain patterns, but suggests an entire other order of being. She writes, poetry

> is not mimesis. Above everything else, it is a formal practice that keeps in view the…otherness of the material medium of language to all that humans do with it—refer, represent, express, narrate, imitate, communicate, think, reason, theorize, philosophize. It offers an experience of another kind of order, a system that operates independently of the production of the meaningful discourse that it enables.[14]

Not only does poetry defamiliarize the communicative function of language, it deliberately highlights the "otherness" of this presumably communicative medium to its mundane function. In forming words, whether into lines, quatrains, or word clouds, the author and reader enter in to "another kind of order," one in which form sets function aside in favor of itself. Form foregrounds the medium as always present and always important to the message; and for Blasing, form is an ontological concern, one which can potentially escape the mundane task of communication altogether. This is why word clouds, as a product of computational criticism, are suitable entry points for scholars of literature; we are accustomed to immersing ourselves in defamiliarized textual objects.

Early modern poets took form just as seriously as Blasing; for them, too, it carried enormous weight. They further saw linguistic form as a way of communicating: form is both a shape given to language, and a kind of content itself. Writing about Shakespeare's position in the early modern literary marketplace, Lamb claims that "words must take a form, even when spoken…words occur *as* and *in* forms, and writers interacted with the market of words *in* and *through* form."[15] Lamb further argues for the primacy of form in the years around the turn of the seventeenth century, in part because of the "historical convergences" of phenomena including the emergence of mercantilism, the book trade, manuscript circulation, and the theater industry, among others.[16] The educational situation in particular contributed to form's importance, since "by the last three decades of the sixteenth century, humanist educational reforms" meant that "multiple generations of a single family…would have received similar instruction in writing production. These students learned to value imitation, copiousness, and rhetorical flexibility."[17] By the time Shakespeare was an active author, the norms of writing copiously and flexibly would have been ingrained not only within a large proportion of his peers, but also across multiple generations of readers who would have received the same training.

As these literary habits became more culturally codified, form became increasingly important as a way of communicating an author's message. Rosalie Colie puts it thus: "Rhetorical education, always a model-following enterprise, increasingly stressed *structures* as well as styles to be imitated in the humane letters—epistles, orations, discourses, dialogues, histories, poems."[18] Not only did authors imitate styles, they began increasingly to imitate structures, or what Colie labels "kinds," of writing. Humanist

education, even as it valued the *copia* of authors and educators like Erasmus, increasingly emphasized the structures given to such copious literary performance.

The thorough entrenchment of humanist educational practices thus enabled what Lamb identifies as a shift in literary taste, endowing form with even more weight. In Lamb's words, "as the sixteenth-century value of copiousness and eloquence gave way to the seventeenth-century value of plainness and directness of expression, writers and readers valued form because it conveyed ideas accurately and appropriately."[19] Form was important not only because of the shape it gave to language, which allowed for communication; form was also important because it carried its own communicative functions. As humanist authors communicated "in and through form," so can modern literary scholars communicate in and through the forms of our scholarship *about* these authors.

George Puttenham's *Arte of English Poesy* (1589) provides one of many contemporaneous guides to the forms and structures that language could take.[20] A masterful compendium of contemporary thinking about the literary and social functions of language, the *Arte* included proliferating definitions of figures including metaphor, simile, synecdoche, hypallage, onomatopoeia, and so on. For Puttenham, each of these figures of speech were a kind of formed language, meant to delight and persuade a reader or hearer on a particular topic on a particular occasion. A striking instance of describing the communicative functions of form is in Puttenham's section on shaped poems; and as we shall see, these functions apply strikingly well to word clouds that I will discuss later. At the end of Book 2, "Of Proportion Poetical," Puttenham, in addition to the figures of speech mentioned above, delineates the "proportion in figure," or the ways in which a poet can

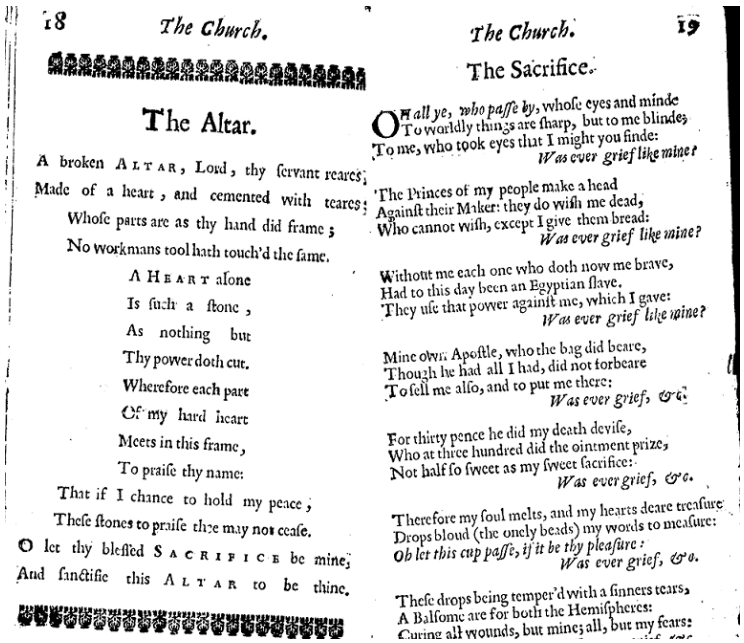
materially shape his verses into shapes such as a lozenge, "romboides," the "triangle," the "square," and the "pilaster, or cillinder."[21] Puttenham elaborates that the "proportion in figure" is

> so called for that it yield an ocular representation, your meters being by good symmetry reduced into certain geometrical figures, whereby the maker is restrained to keep him within his bounds, and showeth not only more art, but serveth also much better for briefness and subtlety of device.[22]

Choosing a physical shape for the verses is premised on the idea that "good symmetry" in meter can be "reduced into certain geometrical figures." When Puttenham advocates for the reduction of a verse, it is because doing so can increase the "briefness and subtlety of device"; to reduce in this context is to distill, in the sense of increased potency.[23] In fact, reducing a verse to a geometric figure is a way for the poet to convey his "art"; when "the maker is restrained" within strict geometrical bounds, he "showeth…more art." In other words, a verse conveys its soundness, and the poet his prowess, by conforming to recognizable geometrical shapes.[24] Form comes first, content second; and even setting aside the difference in genre (e.g., poetry vs scholarship about poetry), modern scholars often prioritize content. For example, we might ask, is a given idea an argument for a blog post? For a journal article? For a book chapter? We often assume, and I think correctly, that the idea must come first and the form second. Word clouds, as computational deformance, shorten the distance between form and content, between idea and written product. They are an argument, in form.

One notable seventeenth-century example of form performing a communicative function is George Herbert's "The Altar." Published posthumously in 1633, "The Altar" is a

devotional poem whose form enacts its content; more precisely, however, the form is part of the content.



18 *The Church.*

The Altar.

A broken ALTAR, Lord, thy ſervant reares;
Made of a heart, and cemented with teares;
Whoſe parts are as thy hand did frame;
No workmans tool hath touch'd the ſame.

A HEART alone
Is ſuch a ſtone,
As nothing but
Thy power doth cut.
Wherefore each part
Of my hard heart
Meets in this frame,
To praiſe thy name:

That if I chance to hold my peace,
Theſe ſtones to praiſe thee may not ceaſe.
O let thy bleſſed SACRIFICE be mine,
And ſanctifie this ALTAR to be thine.

The Church. 19

The Sacrifice.

O all ye, *who paſſe by*, whoſe eyes and minde
To worldly things are ſharp, but to me blinde;
To me, who took eyes that I might you finde:
Was ever grief like mine?

The Princes of my people make a head
Againſt their Maker: they do wiſh me dead,
Who cannot wiſh, except I give them bread:
Was ever grief like mine?

Without me each one who doth now me brave,
Had to this day been an Egyptian ſlave.
They uſe that power againſt me, which I gave:
Was ever grief like mine?

Mine own Apoſtle, who the bag did beare,
Though he had all I had, did not forbeare
To ſell me alſo, and to put me there:
Was ever grief, &c.

For thirty pence he did my death deviſe,
Who at three hundred did the ointment prize,
Not half ſo ſweet as my ſweet ſacrifice:
Was ever grief, &c.

Therefore my ſoul melts, and my hearts deare treaſure
Drops bloud (the onely beads) my words to meaſure:
Oh let this cup paſſe, if it be thy pleaſure:
Was ever grief, &c.

Theſe drops being temper'd with a ſinners tears,
A Balſome are for both the Hemiſpheres:
Curing all wounds, but mine; all, but my fears:

Figure 1: Herbert's The Altar, from The Temple (1633)[25]

A broken A L T A R, Lord, thy servant reares,

Made of a heart, and cemented with teares:

Whose parts are as thy hand did frame;

No workmans tool hath touch'd the same.

A H E A RT alone

Is such a stone,

As nothing but

Thy pow'r doth cut.

Wherefore each part

Of my hard heart

Meets in this frame,

To praise thy Name;

That, if I chance to hold my peace,

These stones to praise thee may not cease.

O let thy blessed S A C R I F I C E be mine,

And sanctifie this A L T A R to be thine.

Herbert presents his heart to God as a "broken altar," which has been "cemented with teares"; he then prays that God's "blessed sacrifice be mine," which will further "sanctifie this altar to be thine." The poem sets itself up as an altar upon which Herbert offers himself; it also takes the form of an altar; it is an altar twice over. In this way, Herbert's altar conforms to Lamb's sense that form "conveyed ideas accurately and appropriately." Herbert offers an altar and/as an altar, in the hope that God will "sanctifie" his offering.

Reading this poem without Puttenham, or perhaps aurally, we lose fully half of the meaning—that meaning which is carried by the form (and would be understood by Herbert's humanist audience). Puttenham describes the altar, or pillar, as the "figure among all the rest of the geometrical most beautiful…By this figure is signified stay, support, rest, state, and magnificence."[26] The altar is the most beautiful of the geometric figures, in part because of what that shape means. The "figure," or form, of the poem "signifie[s]"; form *means* just as much as it *shapes meaning*. Herbert says his altar is "broken," made of the stone which is his heart; that only God has the power to "sanctifie" and thereby heal his heart; that only then will the stones of the altar "praise thee" without "cease."[27] And yet, when Herbert offers this (majestic) form, it is at odds with its humble

content. The form of the poem signals that the poet knows his worth as a poet—the altar is made of a broken heart, *and* it is magnificent.

Shaped poems like Herbert's are certainly heightened examples of form shaping language; but form was as visible to Shakespeare as it was to Herbert. After all, one of the major distinguishing features of poetry—as opposed to prose—is an insistence on the shapes language can take. For Shakespeare, this insistence manifests as repetition. Not only are there repetitive themes within sonnets, themes extend between sonnets so that scholars group poems according to theme; the "young man" and the "dark lady" sonnets are the most obvious example. Add this to the fact that all the sonnets (except two) have fourteen lines, and that many follow the English model of three quatrains and a couplet, and you have a highly-patterned and highly-structured text that returns over and over again to the same kinds of topics, often with the same words and phrases. Booth puts it thus: "as the sonnet form extends the basic verse principle of multiple organization, so Shakespeare's sonnets reflect and magnify the tendencies of the form itself."[28] In other words, the tendency of the sonnet form to multiple organizations of linguistic content are magnified in Shakespeare's repetition of form: the repeated themes of the Petrarchan love sequence are further conveyed in a nearly-unvarying use of form.

In the spirit of the delicate and detailed excavation of meaning typical of much literary criticism, I offer a reading of Sonnet 65, in which Shakespeare negotiates the English sonnet form and the poem's own relation to its literary tradition. In emphasizing the three-quatrains-and-a-couplet structure of a typical sonnet, the poet thereby meditates on the inevitability of temporal destruction, and mourns the fact that even things human think of as durable are subject to decay.

Since brass, nor stone, nor earth, nor boundless sea,
But sad mortality o'ersways their power,
How with this rage shall beauty hold a plea,
Whose action is no stronger than a flower?
O! how shall summer's honey breath hold out,
Against the wrackful siege of battering days,
When rocks impregnable are not so stout,
Nor gates of steel so strong but Time decays?
O fearful meditation! where, alack,
Shall Time's best jewel from Time's chest lie hid?
Or what strong hand can hold his swift foot back?
Or who his spoil of beauty can forbid?
O! none, unless this miracle have might,
That in black ink my love may still shine bright.

The first quatrain asks, if even brass and stone, the earth and the sea, are subject to mortality, what hope can mere beauty have? The next quatrain restates the problem, now intensified with the vocative "O!" at the beginning of line 5; "summer's honey breath" has no hope against the "wrackful siege of battering days," since "rocks" and "gates of steel" will also decay with the passage of time. The sonnet begins to sound despairing, as the problem is restated again and again: if even the biggest and strongest entities can decay, what hope is there for something so perishable as a young man's beauty, "whose action is no stronger than a flower"? The final lines seem to further this hopeless note, as the poet asks, "who [Time's] spoil of beauty can forbid", and then answers "O! none, unless this miracle have

might, / That in black ink my love may still shine bright." All beauty is lost, *unless* the "miracle" of the poet's "black ink" can preserve his love, and implicitly the beautiful beloved.

Such a hope seems tenuous, but is strengthened by the parallel structure of the sonnet's line groupings. Both the second and the third quatrain begin with "O," as does the final couplet. The single-letter vocative carries the form, and in turn, the meaning, since the repetition of the exclamation in the statement of both the problem (even powerful things like rocks and steel will decay, thus the youth's beauty has no hope of survival) and the solution (the miracle of the poetic project) creates an equivalency between problem and solution. While the poet presents the survival of his poems as a "miracle," the linked structure of the sonnet actually strengthens his own "black ink," lending it enough power that it might also be able to compete with the "wrackful siege" of Time's onward march. Rather, Time marches on, but form allows the poet to return again and again to the moment of the vocative "O."

This sonnet worries about several things, chief among them the sheer destructive power of Time. And one of the ways it demonstrates its worry is through repetition, both of words and of ideas. In addition to the repeated exclamation "O" at the second and third quatrains and the final couplet, each quatrain effectively repeats the question, "what hope for frail beauty if even strong things decay with time?" The sonnet is indeed a "fearful meditation," a repetitive mulling-over of a seemingly-unsolvable problem, heightening its musings with repeated words.

The formal division of the sonnet into threes suggests one of the social engagements of the form, namely with the structures of literary tradition. Writing about sonnet 104 ("To

me, fair friend, you never can be old") and its repetitions of the word three, Colin Burrow notes that "sonneteers tended to live life in multiples of three": Burrow cites Horace, Phillipe Desportes, and Pierre Ronsard as sonnet writers who also organize their experience of the beloved in threes. In organizing his "fearful meditation" around three vocative "Os," Shakespeare signals a trope of his poetic tradition.[29] In other words, reinforcing the sonnet's form with a repeated exclamation and thus reiterating the triadic division of the poem reinscribes a trope of the sonnet tradition, solidifying Shakespeare's presentation of his sonnets as precisely a part of this tradition. Content here draws attention to form as the repeated words highlight the quatrains and couplet which structure the poem. And as content points insistently to form, the form reinforces the content, creating a textual object in which the space between form and content shrinks. In this way, this sonnet is very like a word cloud.

An Introduction to Successful Topic Modeling and its Exploratory Productions

Before pursuing the creative deployment of word clouds, I want to offer yet another overview of topic modeling as a computationally-successful procedure; as the old axiom goes, it is important to know the rules in order to break them.[30] In order to successfully topic model, Megan Brett lists four different "things you will need" in order to topic model: "A corpus, preferably a large one"; "Familiarity with the corpus"; "A tool to do the topic modeling"; and "A way to understand your results."[31] In my case, my corpus is Shakespeare's *Sonnets*; my tool was the LDA (latent Dirichlet association) method in R; and in order to understand my results I output the topic models into both .csv format (also known as an Excel spreadsheet) and word clouds. As I ran the LDA algorithm on the *Sonnets*, I performed a few discrete operations on the text. First, all special characters were

stripped out of the text, including numbers and punctuation; each word was then removed from its context, the computational equivalent of throwing every word-token into a bag and shaking it. The corpus became a collection of words, or rather distinct word-tokens (w^1, w^2……w^n), without semantic meaning. The program does not "interpret" the words, nor does it read context; rather, each word is simply a unique token.

This process transforms the *Sonnets* from a structured and ordered text into an undifferentiated collection of word-tokens. The next step is to infer the topics that unite the corpus, or to restructure this collection of word-tokens again, but this time in topics instead of documents. To do so, R takes each word-token from the giant bag and sorts it into an arbitrary number of topics set by the researcher, comparatively poorly at first, and then after several thousand repetitions, with increasing certainty, until (to quote Shawn Graham et al) "it settles on the most likely distribution of words into baskets, which we call topics."[32] For this project, I combined R code from Matthew Jockers and Kailash Awaiti.[33] Awaiti describes the LDA method of topic modeling thusly:

> The basic assumption behind LDA is that each of the documents in a collection consist of a *mixture of collection-wide topics*. However, in reality we observe only documents and words, not topics – the latter are part of the hidden (or *latent*) structure of documents. The aim is to infer the latent topic structure given the words and document. LDA does this by recreating the documents in the corpus by adjusting the relative importance of topics in documents and words in topics iteratively.[34]

LDA assumes that a corpus is made of three things: documents, topics, and words. It also assumes that an individual document in a corpus contains topics that are likely shared

across other documents in that corpus, and that each topic and each document are made up of words. Humans, however, only see two of these three; we only read the words to understand the documents, and tend to miss the shared latent topics which unite the corpus. LDA works backward from a set of documents to guess what kinds of "discourses," or topics, produced those documents.

Topic modeling is thus also a kind of reversal of the creative process: as David Blei puts it,

> the central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as 'reversing' the generative process—what is the hidden structure that likely generated the observed collection?[35]

Given the assumption that a text contains several different latent topics—the *Sonnets* are about love, but they are also about what counts as beauty, and about human procreation, and about literary immortality—topic modeling throws all the words into a bag, then pulls them out again in an attempt to model what those topics look like independent of their literary context.

Scholars have translated the mathematical procedure behind LDA in different ways, most having to do with food. Matt Jockers uses the metaphor of a buffet, from which each author selects the foods/words that they most enjoy;[36] Lisa Marie Rhody describes it as guessing at the produce available at a farmer's market, where each topic is a basket, into which different kinds of produce are sorted by your neighbors, enabling you to deduce the kinds and proportions of produce available at the market.[37] This emphasis on deduction and probability clarify that LDA, in contrast to plug-and-play tools like Wordle, is *not* based

in word counts: so rather than simply observing the produce that made it into the basket (known), topic modeling tries to guess at what the options were to begin with (unknown). Both of these explanations make most sense when the corpus includes multiple authors and multiple documents, including millions and even billions of words: the metaphors of selection, whether from a buffet or a market, represent authorial style as a series of choices that conform to personal preference. Importantly, those choices are not necessarily conscious ones; part of what topic modeling promises is to extract the latent patterns in word usage that characterize different concepts, or topics. The metaphor of choice, then, sees different texts by different authors as, basically, a series of different choices that accumulate toward a single overarching text.

If topic modeling a multiple-author corpus is explained by a series of authorial preferences that build to a larger corpus that can be broken down into its constituent parts (like a buffet or a farmer's market), for a single-author corpus, it is most useful to think of the author's text as a series of combinations of certain set linguistic patterns (like a recipe). As Craig and Arthur Kinney argue, following Stephen Pinker and other cognitive scientists, the brain contains "a recipe or program that can build an unlimited set of sentences out of a finite list of words."[38] A person only knows so many words; in order to produce language (whether creative or communicative), those words are like ingredients which are capable of nearly infinite variation. [39] Further, this set of word-ingredients reveals that human brains contain "a recipe or program" for language use which remains "effectively set throughout an author's career."[40]

Pursuing the idea of language use as a recipe, or a set list of ingredients which are combined and recombined in a variety of ways, we thus might imagine Shakespeare as a

contestant on one of many recent and popular cooking shows. A common challenge is for the host to provide the contestants with a set list of premium ingredients, and then demand they make something delicious that is also frequently counter to the expected outcome. Because of the fixity of the sonnet form, this challenge would instead likely be a list of unusual (and still premium) ingredients, and the contestants would be required to combine them to produce the same dish every time. Let's say that for this show the contestants are asked to make three different versions of eggs benedict, a savory breakfast typically made with English muffins, ham or back bacon, poached eggs, and a hollandaise sauce. The first round would require them to use lobster, grape jelly, and beef stock; the second to make a sweet dessert version with all eggs and eggs replacements forbidden; and the third to incorporate all the ingredients used in the first two versions. Some ingredients are thus repeated; some are omitted; but always, the contestants must produce something recognizable as eggs Benedict.

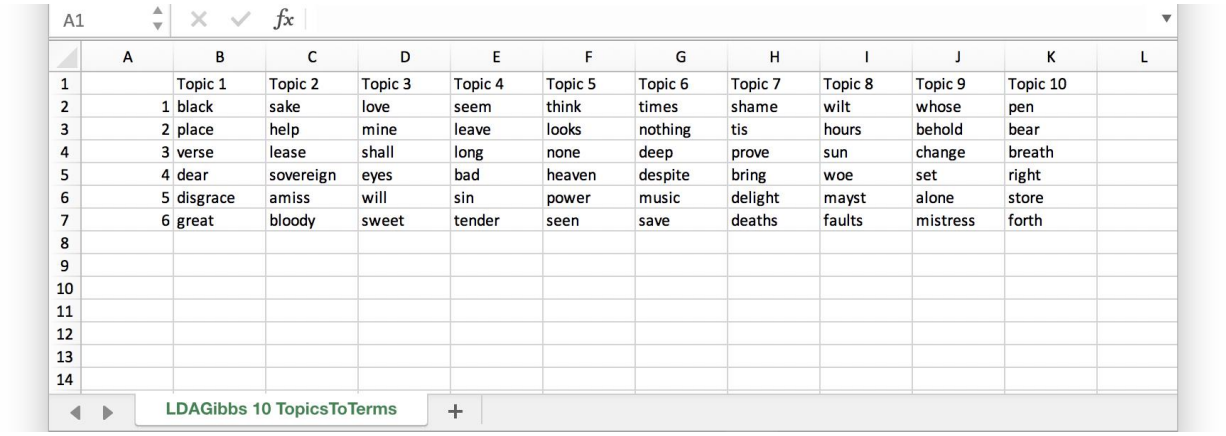
In his *Sonnets*, Shakespeare has a vocabulary of both words and ideas available to him, about love, beauty, mortality, sex, desire, currency, the natural world, and so on; and he combines and recombines them to produce 154 different sonnets. Just as the cooking-show contestants combine, omit, and reuse ingredients, the poet reuses, omits, and combines themes and ideas to produce seemingly-infinite variation in a rigidly-determined form. In Booth's words, the *Sonnets* are structured by "the recurrence in the sequence of themes, situations, ideas, and images"; not only does Shakespeare order his individual poems through repetition, as in Sonnet 65, he also repeats themes and ideas throughout the sequence.[41]

Topic Modeling: Creating Patterns for Analysis

In proposing that topic modeling a poetic corpus can produce poems, I am following an idea suggested by Rhody's article "Topic Modeling and Figurative Language," in which she argues that topic models can indeed be successfully deployed on a poetic corpus precisely because of the interpretive space between topic model and original text.[42] Rhody suggests that topic modeling "can help us to get to new questions and discoveries — not because topic modeling works perfectly, but because poetry causes it to fail in ways that are potentially productive for literary scholars."[43] For Rhody, topic modeling poetry highlights the weak points in topic modeling as a method per se, and allows for a certain kind of "beautiful failure" that nevertheless provides avenues for discovery. Topic modeling, in spite of its bluntness as a tool for excavating meaning, provides the outlines for further fine-detail work. Topic models, and word clouds in particular, thus become both representative and generative of meaning; as I remediate the words in Shakespeare's *Sonnets*, I also create new textual objects that generate their own poetic (because of their formal) function.

Human readers can identify patterns in the *Sonnets*; and with enough time, talent, and dedication, they can identify even patterns that are typically harder to see, as evidenced by Booth's reading of the function words in the sequence. Topic models reveal different kinds of patterns, however. When a corpus is topic modeled, Blei writes that "the interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents."[44] Topics, in other words, are the results of a computational process that attempts to statistically determine what kinds of topics would have come together to make up the document in the first place.[45]

Most topic modeling procedures output the results in tables (typically Excel spreadsheets) such as that in Figure 2.



	A	B	C	D	E	F	G	H	I	J	K	L
1		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	
2	1	black	sake	love	seem	think	times	shame	wilt	whose	pen	
3	2	place	help	mine	leave	looks	nothing	tis	hours	behold	bear	
4	3	verse	lease	shall	long	none	deep	prove	sun	change	breath	
5	4	dear	sovereign	eyes	bad	heaven	despite	bring	woe	set	right	
6	5	disgrace	amiss	will	sin	power	music	delight	mayst	alone	store	
7	6	great	bloody	sweet	tender	seen	save	deaths	faults	mistress	forth	
8												
9												
10												
11												
12												
13												
14												

Figure 2: 10 topics containing the top 6 words in each topic

Figure 2 is the result of asking R to find ten topics, and to only show me the top six words of each topic. (These are not fixed parameters; I could, for example, ask for twenty topics of twenty-five words each.) In these topics, we get an overview of the themes of the sonnets, based on how often the algorithm thinks the words are likely to co-occur. This is in strong distinction from visualization tools like Wordle or Voyant, which only count frequency of occurrence; LDA is a probabilistic method which considers context and proportion (in considering what other words co-occur, and how often). Thus, the top word in each topic is what each topic is most "about"; Topic 3, for example, is most about love, and next-most about possession ("mine"), and so on.

But this spreadsheet is not enough; this is the version of computational scholarship that Ramsay would call simple "fact-checking."[46] For topic modeling to unfold "interpretive possibilities," to meaningfully participate in literary criticism by deepening the discussion around the *Sonnets*, I turn to word clouds. Word clouds are the result of an aesthetic impulse, the desire to see my results in a form that I and other literary critics—readers immersed in the reading of both content *and form*—find more familiar. The truth of this

impulse is demonstrated by the fact that the script needed to output a word cloud does not perform any statistical or mathematical manipulations; instead, word clouds are a remediation of the spreadsheets. I thus argue that in the remediation from spreadsheet to word cloud, we see interpretive deformance at work; as I re-form the words of Shakespeare's *Sonnets*, I enact a process that Puttenham would recognize and of which Ramsay would approve, highlighting textual form as both generative and communicative.

In the following word clouds, I kept the number of topics at ten, and increased the number of word-tokens in each topic to 100. In gaining the increase in context offered by more words, I confirmed that most word clouds seemed to tell us what we already know: they either corresponded fairly directly to particular poems, or corroborated well-known scholarship. Indeed, this supports one of the objections to topic modeling the corpus; human readers have already discovered the themes that make up the structure of the *Sonnets* because of an extensive amount of time and attention devoted to discovering those themes. Here is just one example:

Figure 3: "Praise" word cloud

In Figure 3, the most prominent word is *praise*, surrounded by *memory, pen, pleasure, life,* and *change.* This topic recalls Joel Fineman's argument about the *Sonnets*: they are primarily an iterative epideictic project, in which the poet praises his beloved over and over again, and in so doing writes his own selfhood into being.[47]

Figure 4: "Time" word cloud

Similarly, Figure 4 illustrates Shakespeare's preoccupation with the *summer* of the *youth*'s *beauty*, which the poet's *hand* attempts to preserve against the ravages of *time*; this topic seems to correspond quite nicely to the *Sonnets* themselves, since even conservative investigation reveals eight sonnets that mention summer in conjunction with time and beauty.[48] One such is Sonnet 12:

When I do count the clock that tells the time,
And see the brave day sunk in hideous night;
When I behold the violet past prime,
And sable curls all silver'd o'er with white;
When lofty trees I see barren of leaves
Which erst from heat did canopy the herd,
And summer's green all girded up in sheaves
Borne on the bier with white and bristly beard,
Then of thy beauty do I question make,
That thou among the wastes of time must go,
Since sweets and beauties do themselves forsake
And die as fast as they see others grow;
And nothing 'gainst Time's scythe can make defence
Save breed, to brave him when he takes thee hence.

Addressing this poem to his beloved youth, the poet here observes things that mark the passage of time, such as clocks, violets, grey hair ("sable curls silvered o'er"), trees bare of leaves, and "summer's green" wheat "girded up in sheaves" for the harvest. These observations lead him to question the beauty of the youth, since "sweets and beauties" all

"die as fast as they see others grow." Nothing can guard against "Time's scythe," except to "breed," or bear a son. In this sonnet, part of what is called the procreation sequence, Shakespeare posits a world in which all beauty dies, unless a child can be conceived to carry on the beauty of its father. In the word cloud of Figure 4, these concepts are thrown together by the proximity of key word-tokens.

In the seemingly-precise correspondence between topic and poem, this topic suggests one of the broadest challenges to textual digital humanities scholarship. As Ben Schmidt writes, "The signal criticism of large-scale textual analysis by traditional humanists is that it tells scholars nothing they did not know before."[49] It looks as though we have learned nothing from this project; why put the *Sonnets* into word clouds, if we are sure we know everything about them?

Yet this correspondence is a false one. Rhody puts the dilemma of seemingly-transparent topic models in this way: "Even though the topics appear to have a semantic relationship with the poems because they appear so comprehensible, it is important to remember that semantically evident topics form around a *manner* of speech."[50] Even when we feel like we can simply *read* a word cloud, we must remember that topic models and the word clouds that stem from them are sets of patterns, combined with other co-occurring patterns. This word cloud is not a version of *just* sonnet 12; it is a combination of the eight sonnets (probably more) that examine these themes. Topics are not necessarily connected to the poems they model, nor are they the same as those poems. They are, however, a new "*manner* of speech," a deformance of the original; they are an interpretation, and they also behave like poems because their textual form has pushed against the semantic content they might originally have held.

Shakespeare used the content of Sonnet 65 to emphasize the communicative function of form; following Marshall McLuhan, we thus might say that the poem's medium is its message.[51] The form *is* the content; and the value of topic modeling a highly-canonical corpus is precisely this new context. Topics are not the same as the poems they model; they are, however, remediations of the recipe with which Shakespeare was working. This topic model is not a Shakespearean sonnet: but it is, arguably, a poem, collaboratively written by Shakespeare and me.

Looking at a more opaque word cloud clarifies this point; not all topics will easily confirm existing scholarship.

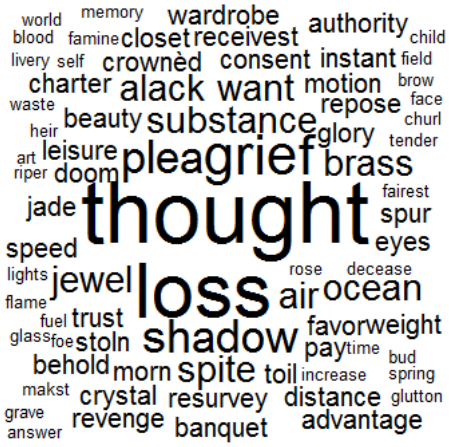


Figure 5: "Thought" word cloud

This topic is dominated by *thought*, and only very slightly less by *loss*; it seems to suggest a topic in which thoughts are the companions of grief and loss, surrounded by substances from the natural world like brass, jewels, oceans, air, shadows, crystal, or jade. Even though there are sonnets which deal with loss and grief (especially in the procreation sequence), this word cloud retains no semantic connection to them. Rather, it exists as its own linguistic object, seemingly self-contained and self-communicative. In its own contained way, this is something that Blasing would recognize as a lyric poem: it is non-referential, in

that its meaning is self-sufficient; it is poetic, in that its words are combined in surprising ways.

Words clouds, then, are not an end or solution. Rather, they are an unfolding of new possibilities to be investigated by scholars, and packaged in a *form* that is familiar and legible to those scholars more than any other. In stripping words from their originary poetic context, the model creates something new; it is an analytical tool when used as a means of exploring a large corpus, but it is also an aesthetic object.[52] The different sizes of each word invite comparison and speculation; in Figure 5, what is the proportional relationship between *spite* and *toil* that the different font sizes identify? *Thought* startles with its dominance, pursued or perhaps accompanied by *loss;* which comes first? More, there are two ways that *thought* works here: if we take *thought* to be an abstract category, the other words clustered around *thought*—grief, plea, shadow, want, alack, spite, memory—are also ephemeral; rather, they speak to absence, to *shadow*, to non-presence. The second point of view is that *thought* should be read as *a* thought, a singular mental event that is thus more concrete. From here, other concrete nouns become more visible: brass, jewel, ocean, substance, jade, air, rose, flame, wardrobe, world, charter. Even the seemingly abstract categories like plea, grief, loss, or shadow take on *a thought's* concrete singularity. Computational processes created new analytical pathways, encouraging new interpretive possibilities and new approaches to one of the most canonical of texts. A surprising topic generates new observations, sometimes about the original text, but often also about its own poetic functioning.

Conclusion

Distant reading's methods and outcomes have been explicitly positioned as opposed to those of close reading. Thus performing close reading on a product of distant reading at first seems counterintuitive. Yet the formal concerns of the early modern period suggest that topic models are a way to reconsider poetic form; thinking of topic models as themselves aesthetic objects opens up further questions about the relationships between humanist studies and the tools, digital or not, with which we perform those studies. Form therefore opens the door to interpretation, to an awareness of the textual object as signaling multiple possibilities. As in Herbert's altar, where the words of the poem say one thing but the form of the poem says another, in word clouds the primary function is to communicate that they are topic models.

A single sonnet, for Shakespeare, could be a statement about how sonneteers lived in threes; 154 of these sonnets, shouting their connections to their social context, became a manifesto. Word clouds of these sonnets are also a manifesto, a striking remediation of words stripped from their original historical, social, and literary contexts that still bellow their own social functions. They thus become, to borrow the words of concrete poet Emmett Williams, a "poetry far beyond paraphrase, a poetry that often asked to be completed or activated by the reader, a poetry of direct presentation—the *word*, not *words, words, words* or expressionistic squiggles—using the semantic, visual and phonetic elements of language as raw materials."[53] In giving poetic language a new shape, in deforming and remediating a collection of mere words, the critic makes choices which are, I argue, parallel to those Shakespeare made when composing his poems. As topic modeling programs break apart and reconstruct their text, a new kind of form emerges: new "models

of language" that do not follow the "laws of their creation." In putting the *Sonnets* through an algorithm, and then seeing them anew in word clouds, computational scholars also participate in a (de)formalizing linguistic practice which produced the very objects of our study in the first place; we too shape language in ways that are not directly communicative. In other words, topic models behave like poems, in that they ask us to reconsider what, and more importantly how, we think we know things about language. In Rhody's words, topic modeling figurative language

> means allowing for a…beautiful failure — not a failure of language, but a necessary inclination toward form that involves a diminishing of language's possible meanings. However, the necessarily reductive methodology of sorting poetic language into relatively stable categories, as topic modeling suggests, yields precisely the kind of results that literary scholars might hope for — models of language that, having taken form, are at the same moment at odds with the laws of their creation.[54]

In making anew the shapes of already-poetic language, and especially when doing so in a way which is counter to the communicative language on which topic model algorithms are trained, a new and "beautiful" form emerges: new "models of language" that do not follow the "laws of their creation." To return to Blasing's definition of a lyric poem, topic models are also "a formal practice that keeps in view the…otherness of the material medium of language to all that humans do with it."[55] In putting the *Sonnets* through an algorithm, and then seeing them anew in word clouds, we also participate in a formal linguistic practice; we too shape language in ways that are not directly communicative.

And yet word clouds *do* communicate; and I argue both that their medium is their message, and that they address what Alan Liu calls "the paradox of 'notation versus

presentation.'"[56] For Liu, who substantively adds to McLuhan's arguments about new media, "notation" is when the audience (whether reading or viewing or hearing) experiences "new media as a kind of script or programming language that must be mentally decoded[, whereas] 'Presentation' means rendering the script intuitively or immediately."[57] The audience either puzzles through the new media object piece by piece, or immediately *gets it*; Liu asks, "how will we ever change from clumsy searching, sampling, and text- or data-mining to directly grokking the data patterns" of the new media we encounter?

I propose that word clouds exist at the exact center of this paradox, and in fact serve as an object that allows the reader to experience both its terms at once. As a formal object, word clouds suggest both an immediate, impressionistic "grokking" of the underlying "data patterns" of the thing they remediate, and they invite the reader to perform the searching, delicate, sometimes-clumsy work of meaning-making that is close reading. And this is the value of word clouds, and, I would argue, of the products of computational humanities scholarship more broadly: such products exist *both* as an interpretation of an originary creative text, *and* as their own creative objects that provoke further interpretation. Word clouds are notation *and* presentation, medium *and* message, form *and* content: and in being both at once, they step outside of the paradox and become, like the *Sonnets*, the site of simultaneous mediation and provocation.

[1] Lisa Samuels and Jerome McGann, "Deformance and Interpretation," *New Literary History* 30, no. 1 (1999): 25.

[2] Samuels and McGann, 25.

[3] Samuels and McGann, 40.

[4] Samuels and McGann, 34.

[5] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Chicago: University of Illinois Press, 2011), 8.

[6] Ramsay, 9.

[7] Ramsay, 10.

[8] Stephen Booth, *Essay on Shakespeare's Sonnets* (New Haven: Yale University Press, 1969). See especially pgs. 12, 26-27, 35, 60, 82-83.

[9] Douglas Duhaime topic models the presence of characters onstage in various genres in "Classifing Shakespearean Drama with Sparse Feature Sets," accessed August 1, 2018, http://douglasduhaime.com/posts/classifying-shakespearean-drama-with-sparse-feature-sets.html; see also Eric T Nalisnick and Henry S Baird, "Character-to-Character Sentiment Analysis in Shakespeare's Plays," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, 479–83. Laura Estill and Meneses, "Is Falstaff Falstaff?" *Digital Studies/Le Champ Numerique* 8.1 (23 Jan 2018). http://www.digitalstudies.org/articles/10.16995/dscn.295/

[10] Estill and Meneses, "Is Falstaff Falstaff?"

[11] Hugh Craig, "Shakespeare's Vocabulary: Myth and Reality," *Shakespeare Quarterly* 62.1 (2011): 68.

[12] To add yet another obstacle to topic modeling the sonnets, natural language processing tools (including topic modeling algorithms) were developed on, and most successfully deployed for, corpora of prose, both fictional and nonfictional, written mostly from the mid nineteenth century onward. This means that topic modeling algorithms can have a hard time working with texts before 1800 or so because of the variance in spelling and speech patterns. See David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 1, 2012): 77, https://doi.org/10.1145/2133806.2133826.

[13] Blei, 77–78.

[14] Mutlu Konuk Blasing, *Lyric Poetry: the Pain and the Pleasure of Words* (Princeton: Princeton University Press, 2008). 2.

[15] Jonathan P. Lamb, *Shakespeare in the Marketplace of Words* (Cambridge: Cambridge University Press, 2017), https://doi.org/10.1017/9781108147729. 22

[16] Lamb, 19–21.

[17] Lamb, 19–20.

[18] Rosalie Colie, *The Resources of Kind: Genre-Theory in the Renaissance* (Berkeley: University of California Press, 1973), 4.

[19] Lamb, *Shakespeare in the Marketplace of Words*, 22.

[20] George Puttenham, *The Art of English Poesy*, ed. Frank Whigham and Wayne A. Rebhorn, Cornell Paperbacks (Ithaca: Cornell University Press, 2007).

[21] Puttenham, 181.

[22] Puttenham, 179.

[23] *Reduce* carries both the sense of "reduce in size," but also "increase in potency"; both are at play here. "Reduce, V.," in *OED Online* (Oxford University Press), accessed August 10, 2018, http://www.oed.com.ezproxy.amherst.edu/view/Entry/160503.

[24] For a thorough discussion of Puttenham's shapes in the *Arte,* see chapter 2 of Miriam Jacobson, *Barbarous Antiquity: Reorienting the Past in the Poetry of Early Modern England* (Philadelphia: University of Pennsylvania Press, 2014).

[25] George Herbert, *The Temple Sacred Poems and Private Ejaculations. By Mr. George Herbert, Late Oratour of the Universitie of Cambridge.*, Early English Books, 1475-1640 / 1990:04 ; Early English Books, 1475-1640 / 2078:06 ([Cambridge] : Printed by T. Buck, and R. Daniel, printers to the Universitie of Cambridge, 1633., 1633).

[26] Puttenham, *The Art of English Poesy*, 186.

[27] For more on the tension between self and God in Herbert's poetry, see Michael Schoenfeldt, 'George Herbert, God, and King', in *Early Modern English Poetry: A Critical Companion*, ed. by Patrick Cheney (New York: Oxford University Press, 2007), 264-277.

[28] Stephen Booth, *Essay on Shakespeare's Sonnets* (New Haven: Yale University Press, 1969), 174.

[29] The most famous discussion of the relationship of the poems to their literary tradition (specifically their belated or outmoded status, is still Joel Fineman, *Shakespeare's Perjured Eye: The Invention of Poetic Subjectivity in the Sonnets* (Berkeley: University of California Press, 1986). For more, see Emily Vasiliauskas, "The Outmodedness of Shakespeare's Sonnets," *ELH* 82, no. 3 (2015): 760.

[30] Rhody provides one of the better introductions to the principles behind topic modeling I have read; there are several others by scholars like Scott Weingart, Ted Underwood, and

Matt Jockers. Matthew Jockers, "The LDA Buffet Is Now Open; or, Latent Dirichlet Allocation for English Majors Matthew L. Jockers," accessed August 24, 2018, http://www.matthewjockers.net/2011/09/29/the-LDA-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/. Scott Weingart, "Topic Modeling for Humanists: A Guided Tour," the scottbot irregular, July 25, 2012, http://www.scottbot.net/HIAL/?p=19113. Ted Underwood, "Topic Modeling Made Just Simple Enough.," *The Stone and the Shell* (blog), April 7, 2012, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/. For a basic overview on downloading and installing R, I recommend Taryn Dewar, "R Basics with Tabular Data," *Programming Historian*, September 5, 2016, https://programminghistorian.org/en/lessons/r-basics-with-tabular-data.

[31] Megan R. Brett, "Topic Modeling: A Basic Introduction," Journal of Digital Humanities, April 8, 2013, http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.

[32] Shawn Graham, Scott Weingart, and Ian Milligan, "Getting Started with Topic Modeling and MALLET," *Programming Historian*, September 2, 2012, https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

[33] More specifically, I modified Awaiti's code to clean and preprocess the text, used Jocker's code for topic modeling and outputting word clouds of an XML file from the Folger Library, and then modified Jocker's word cloud code with the color packages from Awaiti's site.

[34] Kailash Awaiti, "A Gentle Introduction to Topic Modeling Using R," *Eight to Late* (blog), September 29, 2015, https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/.

[35] Blei, "Probabilistic Topic Models," 79.

[36] Jockers, "The LDA Buffet Is Now Open; or, Latent Dirichlet Allocation for English Majors Matthew L. Jockers."

[37] Lisa M. Rhody, "Topic Modeling and Figurative Language," Journal of Digital Humanities, April 7, 2013, http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

[38] Pinker, qtd in D. H Craig and Arthur F Kinney, *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge: Cambridge University Press, 2012), 4–5.

[39] Ibid., 209. For a more exhaustive computational examination of early modern linguistic patterns, see Jonathan Culpepper and Merla Kytö, *Early Modern English Dialogues: Spoken Interaction as Writing* (Cambridge: Cambridge University Press, 2010). Craig and Kinney are specifically working to parse authorship of several plays typically thought to be written by Shakespeare in collaboration with other playwrights, though their conclusions apply more broadly. By analyzing patterns of speech which recur in works known to be by Shakespeare, they are able to identify words and phrases that Shakespeare seemed to favor; they thus tentatively assign authorship in coauthored plays. These authorship debates are far from settled, however: the SHAKSPER listserv continues a months-long (as of this writing in December 2018) and heated debate about methods, conclusions, and replicability. "SHAKSPER Current Postings," accessed December 18, 2018, https://www.shaksper.net/current-postings. Pervez Rizvi has recently published a critique of Craig and Kinney's methods, and subsequently questions their results. Pervez Rizvi, "The Interpretation of Zeta Test Results," *Digital Scholarship in the Humanities*, accessed December 20, 2018, https://doi.org/10.1093/llc/fqy038.

[40] Craig and Kinney, 4, 209. Craig and Kinney's conclusions are applicable to authors who remain neurotypical with age. In "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists," Xuan Le et al study the corpora of Agatha Christie, P.D. James, and Iris Murdoch. These authors were suspected of, died of, and remained free of Alzheimer's disease respectively. Le et al find that while Christi and Murdoch both demonstrated "A major loss in vocabulary (revealed by type/token ratio and word-type introduction rate), an increase in repetition of fixed phrases and of content words within close distance, a deficit in noun tokens and a compensation in verb tokens, and a pronounced increase in fillers" (457), Murdoch's lexicon remained relatively stable. Thus, a healthy aging brain retains the linguistic capacity and preference of youth. Xuan Le et al., "Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of Three British Novelists," *Literary and Linguistic Computing* 26, no. 4 (December 1, 2011): 435–61, https://doi.org/10.1093/llc/fqr013.

[41] Booth, *Essay on Shakespeare's Sonnets*.

[42] Rhody, "Topic Modeling and Figurative Language."

[43] Rhody.

[44] Blei, "Probabilistic Topic Models," 79.

[45] Blei does point out that "calling these models 'topic models' is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA."

The results look like topics because of how humans process language, and how texts are typically composed.

[46] Ramsay, *Algorithmic Criticism,* 10.

[47] Fineman, *Shakespeare's Perjured Eye: The Invention of Poetic Subjectivity in the Sonnets*.

[48] A search in "Open Source Shakespeare" reveals that 5, 6, 12, 13, 54, 65, 68, and 104 share references to summer and to time and/or beauty. "Open Source Shakespeare," accessed July 26, 2018, https://www.opensourceshakespeare.org/views/sonnets/sonnet_view.php?Sonnet=all&pleasewait=1&msg=pl.

[49] Ben Schmidt, "Do Digital Humanists Need to Understand Algorithms?," ed. Lauren Klein, and Matt Gold, Debates in the Digital Humanities, 2016, http://dhdebates.gc.cuny.edu/debates/text/99. Accessed August 1, 2018.

[50] Rhody, "Topic Modeling and Figurative Language."

[51] Marshall McLuhan, *Understanding Media: The Extensions of Man*, 1st MIT Press ed (Cambridge, Mass: MIT Press, 1994), 7.

[52] I have not addressed here other aesthetic decisions that can be made in coding word clouds. For example, I can change values to rotate words, to add color, to increase or decrease the number of words included, and so on. Each of those decisions affects interpretation: humans are very good at seeing connections among similar things, and thus colors can affect the way the topic model is interpreted. While changes in interpretation based on aesthetics might affect the statistical validity of the word cloud, here they are precisely the point. After all, much has been made in literary studies of the practice of close reading, and the changes in meaning that result from a different verb, a crossed-out

adjective, an added comma. Puttenham, in a striking example from a time when orthography remained fairly flexible, even argues that changing the spelling of a word affects its pronunciation; the poet should therefore think about how he wishes to spell words, since that will affect how readers understand his poems.

[53] Emmett Williams, *An Anthology of Concrete Poetry* (New York: Primary Information, 2013), vi. For more on concrete poetry, see Jon M. Tolman, "The Context of a Vanguard: Toward a Definition of Concrete Poetry," *Poetics Today* 3, no. 3 (1982): 149, https://doi.org/10.2307/1772395; Decio Pignatari and Jon M. Tolman, "Concrete Poetry: A Brief Structural-Historical Guideline," *Poetics Today* 3, no. 3 (1982): 189, https://doi.org/10.2307/1772398.

[54] Rhody, "Topic Modeling and Figurative Language."

[55] Blasing, *Lyric Poetry the Pain and the Pleasure of Words*, 2.

[56] Alan Liu, "Imagining the New Media Encounter," in *A Companion to Digital Literary Studies*, 2007, http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-3-1&toc.id=0&brand=9781405148641_brand.

[57] Liu.